25th ANNUAL NATIONAL CONFERENCE ON MANAGING ENVIRONMENTAL QUALITY SYSTEMS

APRIL 24-27, 2006

Marriott Renaissance, Austin, Texas

Technical Papers

Advances in Sampling and Monitoring

- B. Pulsipher,Visual Sample Plan - 8:30 AM
- D.Crumbling, Heterogeneity Rules: Implications for Environmental Statistics - 9:30 AM

New Visual Sample Plan (VSP) Module Pertinent to Environmental Trend Monitoring

Brent A. Pulsipher, John E. Wilson and Richard O. Gilbert
Pacific Northwest National Laboratory
Richland, WA 99352

Abstract

25th Annual Conference on Managing Environmental Quality Systems
Austin, TX, April 24-27, 2006

Visual Sample Plan (VSP) is a Data Quality Objectives (DQO) based software tool that helps determine the required number and optimal placement of samples and performs statistical tests on the resulting data to support confident decisions. A new module is being added to better address long-term monitoring needs. This module is focused on trend detection and estimation. The objective of many long-term monitoring programs is to ensure that sampling is adequate to ensure a timely and likely detection of a significant trend. VSP allows the user to specify the size of the change that must be detected, the tolerable decision error rates, and the rate of sampling, then determines the number of samples required at each sampling interval.

The VSP trend detection module also permits analysis of the data with and without seasonal adjustments. When seasonality effects are present, the Seasonal Kendall trend test is performed whereas if no seasonality effects exist, the Mann-Kendall test is performed. Least squares linear regression fits, nonparametric linear fits of the trends, and normality tests and descriptive statistics on the data and residuals are also provided in VSP. This paper will illustrate these methods.

Introduction

Visual Sample Plan (VSP) is a Data Quality Objectives (DQO) based, user-friendly, highly visual software tool to help determine the required number and optimal placement of environmental samples, conduct data quality assessments, and compute statistical tests to support confident decisions. This paper focuses on a new module on trend detection, estimation and data quality assessment that is being added to VSP to better address the needs of many long-term monitoring programs for environmental media.

The set of statistical tools that are currently available in VSP (Version 4.5) to detect, estimate and describe trends at monitoring stations are:

- <u>a simulation routine to compute the number of samples</u> required by the nonparametric, distribution-free Mann-Kendall (MK) and Seasonal-Kendall (SK) tests (Helsel and Hirsch 1992, Chapter 12; Gilbert 1987, Chapters 16 and 17;

Hirsch, Slack and Smith 1982) to detect upward or downward monotonic trends of importance.

- data quality assessment analyses: descriptive statistics for data collected over time, plots of trend data and residuals (histograms, box-and-whisker plots, Q-Q plots and linear regression plots), and tests for normality of residuals from linear regression using the Shapiro-Wilk and Lilliefors tests (Gilbert 1987, p. 158; Lilliefors 1967, 1969).
- statistical testing for monotonic trends using the MK and SK tests. Note that the SK test takes into account seasonal variation by computing the MK test on each of the seasons separately, and then combining the results. The SK test is appropriate when there is a single pattern of trend across all seasons (Helsel and Hirsch, 1992, p. 344).
- statistical estimation of trends using Sen's nonparametric estimator of trend with confidence intervals (Sen 1968, Gilbert 1987, p. 217), nonparametric linear fits of trend (Helsel and Hirsch, 1992, p. 266), and least squares linear regression (trend) estimation.

VSP grew out of the development of the ELIPGRID-PC code at ORNL (Davidson 1995), which determines the number and placement of samples on a systematic grid needed to detect hot spots of specified circular or elliptical size and shape with specified probability. ORNL and the Pacific Northwest National Laboratory (PNNL) joined forces in 1997 to produce VSP versions of increasing size and capability, the latest being Version 4.5. VSP and its user's guide for Version 4.0 (Hassig, et al., 1995) are available for free download from http://dqo.pnl.gov/vsp.

Version 4.5 contains modules for many decision objectives including estimating means and confidence intervals, comparing averages, proportions or individual measurements to threshold values, comparing two sites or a site to a reference area, finding hot spots or unexploded ordnance (UXO) target areas, accessing the success of UXO cleanup using statistical acceptance or compliance sampling, delineating boundaries of contamination, sampling buildings to assess presence and degree of biological/chemical./radiological contamination, and testing for trends over time. Version 4.5 of VSP supports the following sampling design strategies: judgmental, simple random, systematic (grid), stratified, collaborative, ranked set, adaptive cluster, and sequential over time. VSP development has been funded by DOE (EH-3), EPA (OSWER and OEI), DoD (SERDP and ESTCP), DHS (TSWG) and the Atomic Weapons Establishment (AWE) in England and Wales.

Example of Using the VSP Trend Module

The trend module in VSP is accessed by clicking **Sampling Goals** >> **Detect a Trend**>>**Data Not Required to Be Normally Distributed**, followed by either **No Seasonality** or **Seasonality,** which brings up the dialog box for the MK test or the SK test, respectively. The dialog box that appears for the SK test is shown in Figure 1.0. The dialog box for the MK test is the same except there is no **season** tab at the top of the box. The VSP user inputs the information into the dialog box, which VSP uses to compute the number of samples needed over time. The dialog box inputs are:

- the alternative hypothesis: ‘upward trend” or “downward trend” (the null hypothesis is always “no trend”)
- the allowed probability (alpha) the MK test will falsely reject the null hypothesis
- the allowed probability (beta) the MK test will falsely accept the null hypothesis
- the change per time period (linear trend) that is important to detect
- the desired time interval between sampling events
- the anticipated standard deviation of the residuals for the assumed linear regression trend (based on past data)
- the threshold value (action level) that if crossed by the trend line triggers a specified action

Clicking the **season** tab at the top of the dialog box brings up the season dialog box shown in Figure 2.0, which allows the user to define the season. The season options are: month, week, hour, calendar quarter (spring, summer, fall, winter), or a user-defined month, week or hour season. Figure 2.0 shows two seasons: Jan-June and July-Dec.

After measurements are obtained they are entered into the **Data Entry** box in VSP that is accessed by clicking the **Data Analysis** tab. Figure 3.0 shows hypothetical data obtained approximately every 6 months between 2001 and 2004. Note that two measurements were made for season 1 in 2001. VSP uses the median of those two values in all computations; similarly for the two data obtained in season 2 in 2002.

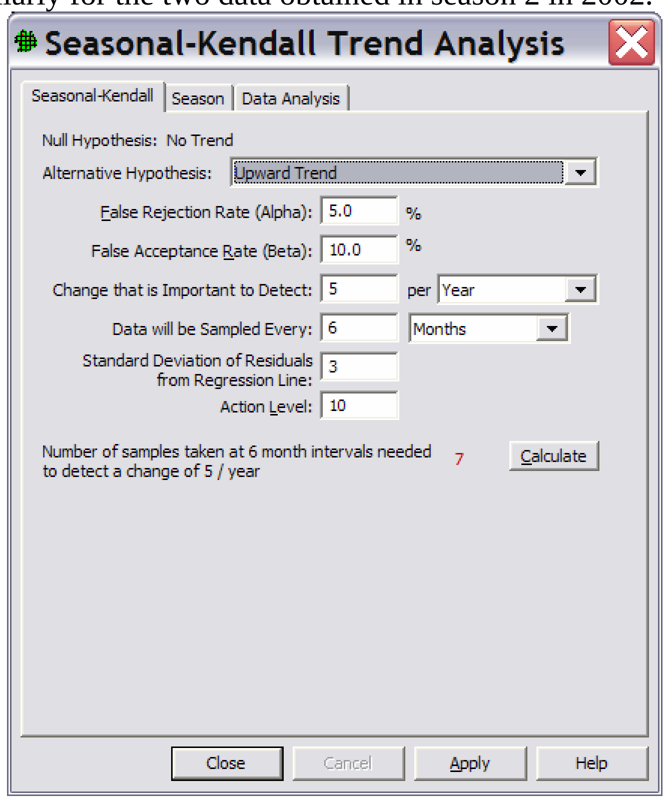

Figure 1.0 VSP dialog box for the SK test. The default values shown are replaced by the VSP user’s values.

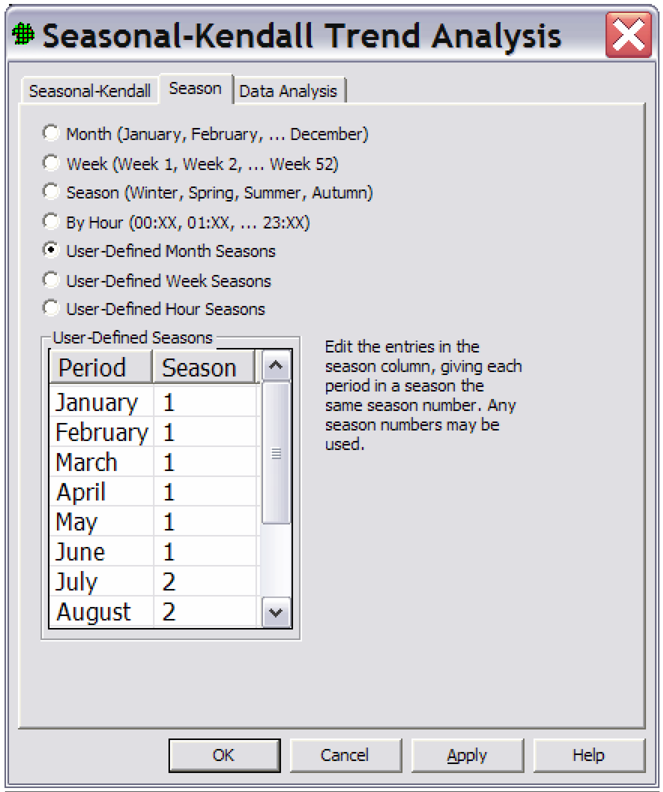

Figure 2.0 Season dialog box for the SK test.

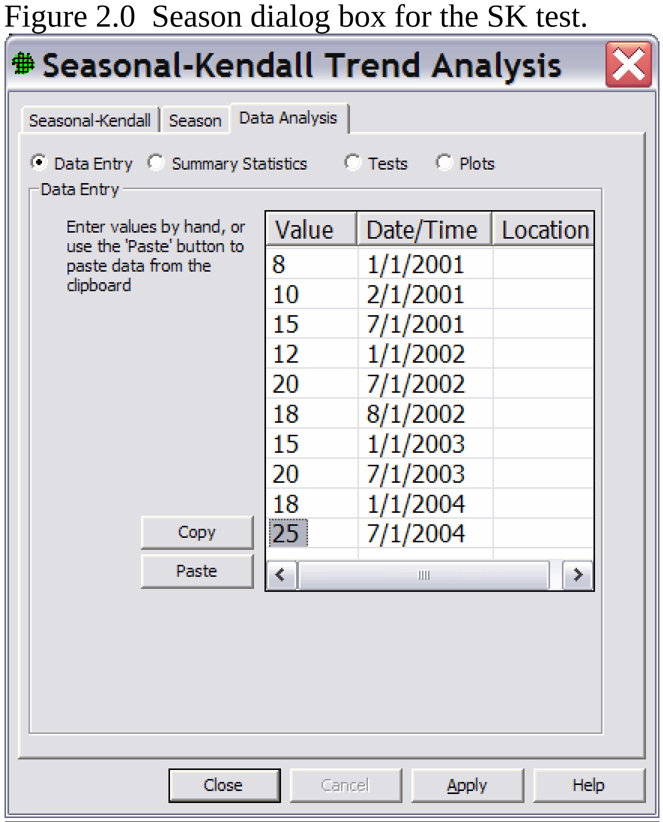

Figure 3.0 Data Entry box for the SK test showing hypothetical data

Figure 4.0 shows the time plot of the hypothetical data in Figure 3.0. Figure 5.0 shows the statistical test results obtained by VSP for the data in Figure 3.0. Starting from the top of the page are the nonparametric linear line (Y = 11.1 + 3.5X), Sen's nonparametric estimate of slope (3.5 units per year) with it's 95% confidence interval (2.2 to 5.8 units per year), the estimated least squares linear regression line displayed in Figure 5.0 (Y = 10.3 + 3.7X), the time when the linear line crosses the threshold of 10 units (12/2/2000), and the results of the SK test (indicates an upward trend was present at the 5% significance level). Clicking the **Summary Statistics** button on the **Data Analysis** tab will display the descriptive statistics for the data set (including minimum, maximum, mean, median, standard deviation, etc) and also for the residuals from the linear regression and the results of the test for normal distribution of the data and of the residuals.

Future Work on VSP Trend Module

Some additional statistical tools being considered for inclusion in VSP trend module include:

- t-test for zero linear slope (trend) of a linear regression line assuming residuals from the line are normally distributed with a constant variance over time.
- Hodges-Lehmann estimator of step-trend (sudden change) as an alternative to two-sample t test (Helsel and Hirsch, 1992, p. 348).

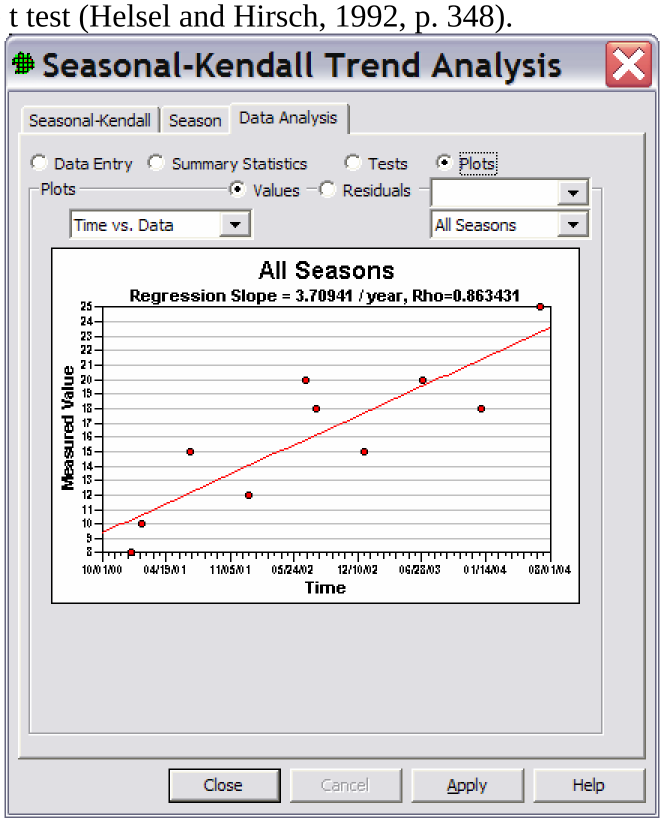

Figure 4.0. Time plots of the hypothetical data in Figure 3.0

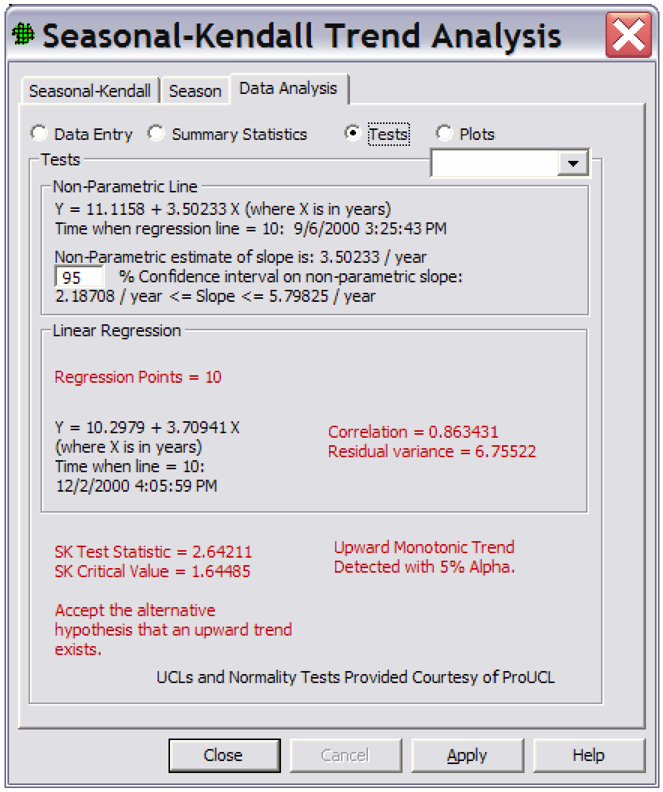

Figure 5.0. VSP results of statistical tests based on the data in Figures 3.0 and 4.0

- testing for homogeneous seasonal trends, i.e., testing to assess if the seasons are behaving in a similar fashion (an assumption of the SK test) (Gilbert 1987, Helsel and Hirsch 1992, p. 345).
- a trend data smoothing method such as LOWESS (LOcally WEighted Scatterplot Smooth) (Helsel and Hirsch 1992, p. 334).
- statistical methods for handling non-detects in trend data (Helsel 2005).
- change and outlier detection tests

References

Davidson, J.R. 1995. *ELIPGRID-PC: Upgraded Version.* ORNL/TM-13103, Oak Ridge National Laboratory, Oak Ridge, TN.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*, Wiley, NY.

Hassig, N.L., J.E. Wilson, R.O. Gilbert, B.A. Pulsipher, and L.L. Nuffer. 2005. *Visual Sample Plan Version 4.0 User's Guide*, PNNL-15247, Pacific Northwest National Laboratory, Richland, WA, July 2005.

Helsel, D.R. 2005. *Nondetects and Data Analysis, Statistics for Censored Environmental Data*, Wiley, NY.

Helsel, D.R. and R.M. Hirsch. 1992. *Statistical Methods in Water Resources,* Elsevier, NY.

Hirsch, R.M., J.R. Slack, and R.A. Smith. 1982. "Techniques of trend analysis for monthly water quality data," Water *Resources Research* 18(1):107-121.

Lilliefors, H.W. 1967. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association* 62:399-402.

Lilliefors, H.W. 1969. "Correction to the paper `On the Kolmogorov-Smirnov test for normality with mean and variance unknown`," *Journal of the American Statistical Association* 64:1702.

Sen, P.K. 1968. "Estimates of the regression coefficient based on Kendall's tau," *Journal of the American Statistical Association* 63:1379-1389.

Heterogeneity Rules: Implications for Environmental Statistics

Deana M. Crumbling, Michael L Adam, Stephen Dyment
US EPA Office of Superfund and Technology Innovation Program
1200 Penna. Ave., NW MC 5102G, Washington, DC 20460

Classical statistics are recommended as a means to evaluate environmental conditions and contamination. However, a primary assumption of the classical statistical model is that the population attribute is randomly distributed. Yet, one of the hallmarks of environmental conditions is heterogeneity at scales which influence how data should be generated and interpreted. Environmental heterogeneity arises because physical and biological mechanisms influencing most environmental conditions do not function randomly; most operate through cause-and-effect. Erroneous conclusions are likely when data collection has not be matched to the spatial and temporal scales of heterogeneity within the context of the hypothesis to be tested or decision to be made. Those drivers are critical to selecting the appropriate statistical mode used to structure data generation and interpretation. This is especially true within certain specialized environmental activities, such as characterizing and remediating contaminated sites. Improperly applying classical statistical models in site cleanup can cause erroneous cleanup decisions. Inefficient cleanups may have severe ramifications for human and ecosystem health, local economies, local and national perceptions of governmental effectiveness, agency budgets and political capital. Since cleanup programs are an importance part of EPA's mission, this paper's discussion will focus on the roles and pitfalls of statistics in site cleanup. However, these lessons are relevant to other environmental studies and regulatory programs.

Introduction

It is well known that the output of inappropriately applied statistics may produce misleading conclusions. This is especially true when dealing with sites contaminated with hazardous chemicals with their inherent heterogeneity. The physical mechanisms governing contaminant release, deposition, migration/transportation, sorption, and degradation create positive and negative spatial correlations for contaminants throughout the matrix. Basic assumptions of the classical statistics model—random distribution and independence among data points within the targeted population—are easily violated when project design fails to consider the specific site decision-making needs. Special precautions are required to ensure correct conclusions when site realities and statistical models are mismatched.

Several important concepts must be considered and reconciled when selecting and tailoring statistics to waste site cleanup scenarios. These concepts include

- the conceptual site model (CSM) and contaminant populations;
- spatial scales of heterogeneity;
- sample support and its relationship to concentration variability; and
- decision-driven populations.

The balance of this paper will introduce these concepts and their importance for proper application of statistics and efficient contamination cleanup. A more detailed discussion will be presented in the accompanying workshop, "Avoiding Statistical Pitfalls in Environmental Science."

The Conceptual Site Model and Contaminant Populations

A conceptual site model (CSM, http://www.triadcentral.org/mgmt/splan/sitemodel/) encompasses everything that is currently known about the site and what needs to be known so that site issues can be addressed. The CSM is constructed from historical information related to site use and contaminant release; information about matrix composition that is pertinent to contaminant fate and migration; regulatory and community concerns about the site and its future; and scientific and engineering judgments about risks posed by the site and anticipated risk mitigation measures. Sometimes CSMs can be limited to a single description, graphic, or modeled representation. But often multiple diagrams or modeling tools are needed to portray various aspects of contaminant behavior in its complex environmental setting. A preliminary CSM is developed at the beginning of the project from existing information. The CSM will be tested, modified and refined as more information and data are obtained throughout the site's lifecycle. The model is mature when it can be used to make confident predictions about exposure risk and cost-effective risk reduction strategies. CSM refinement necessarily continues throughout the lifecycle since contaminants may continue to move and/or degrade under the influence of natural conditions or active remediation.

It is crucial to develop a preliminary CSM before data collection begins. A "draft" CSM postulates contaminant presence and mass, movement, exposure pathways as a function of site reuse, and options for reducing risk. Different contaminant populations are created by differential impacts to environmental media as a result of contaminant release and migration. The preliminary CSM serves as a testable hypothesis of the presence and distribution of contaminant populations. This hypothesis will be confirmed or rejected and modified as evidence accumulates. The preliminary CSM allows information gaps to be identified, and suggests the variables that must be controlled if interpretable data are to be generated.

Statistical implications: The CSM is key to defining the statistical populations relevant to project decisions, the appropriate statistical model, and the statistical parameter(s) of interest. Without a CSM that captures contaminant distributions in the context of intended site decisions, data representative of the hypotheses to be tested and the decisions to be made cannot be generated. Statistics users must be aware that the classical statistical model may not hold for contaminated sites. Classical statistics rests on the assumptions that sample results are taken from a single population. Samples also must be "independent," that is, they do not show correlation where sample results are related to each other. When those assumptions hold, random sampling can adequately and fairly characterize the population (Steel et al, 1997).

Spatial Scales of Heterogeneity

The sporadic, localized, and mechanism-specific nature of most contaminant releases creates spatial and temporal heterogeneity. Interaction with environmental matrix components after release typically enhances heterogeneity even further. Environmental matrices—soil, surface water systems and associated sediments, aquifers (containing groundwater)—are inherently heterogeneous in composition. Contaminant interactions at the bulk and molecular scales with matrix components create contaminant concentrations that vary continuously over scales ranging from tiny (millimeter and smaller) to small (inches to feet) to medium (feet to yards) to large (yards to miles) dimensions.

- **Although perhaps homogeneous in appearance, surface water systems differ in attributes such as pH, dissolved mineral content, suspended particles, microbial communities, and flow regimes. Water properties can change even within a "single" water body based on in-water structures, depth, influents, temperature gradients, flow currents, vegetation, and other watershed variables.**
- **Soil, sediments and aquifer materials are composed of a myriad of different minerals, particle sizes, organic substances, biota and other materials that interact with introduced chemicals according to the chemical and physical properties of both matrix and contaminant.**

Few people recognize that data results are highly dependent on how the sample is collected and processed. Contaminants bind preferentially to certain particles and particle sizes, which creates a "nugget effect" (ITRC, 2003a). Since concentration is a function of both analyte mass and matrix volume, the same mass of contaminant in a "nugget" will produce different concentration results depending on how much "clean" matrix is processed along with the "nugget." This effect is very common for contaminants. Extrapolating concentration results of tiny analytical samples to tremendously larger volumes of matrix (the expectation for a "representative" sample) is highly uncertain and potentially misleading when sampling variables are not controlled.

<u>Statistical implications:</u> Sampling environmental matrices is entirely different from sampling the types of populations for which statistics was developed. There is no discrete, individual, natural sampling unit for soils or water. Sampling units are currently chosen out of convenience or precedent, not out of scientific or statistical consideratison for controlling sampling variability. This has tremendous implications for the interpretation of data, implications that cannot be controlled after the fact through statistical calculations and software.

Sample Support and Its Relationship to Concentration Variability

As introduced in the previous section, sample/data representativeness requires conscious control over the variables that influence concentration results. Otherwise, extrapolation of results beyond the sample aliquot actually extracted or digested for analysis is invalid. But what determines the proper way to control those variables? Should the variables be controlled so that the result is an average over some bulk volume? Or should a proportion other than average be used? How is the volume of sample selected? Should compositing be used? If so, how? How should data results be compared to regulatory thresholds?

Answering these questions invokes the concept of "sample support" a term that has appeared in EPA guidance documents at least since 1993 (USEPA, 1993). Sample support refers to the physical dimensions of a sample that influence analytical results. This includes the mass of the sample and the three spatial dimensions of a sample's height, depth and width as it is extracted from its parent matrix. This concept applies both to water and solid samples. As noted above, particle size can also be a highly influential physical parameter influencing analytical results. Particle size effects have been shown to cause lead concentrations to increase over 200X as the particle size fraction went from larger to smaller (ITRC, 2003b). Assuming no attempt at appropriate matrix homogenization, data results depend on what particle size happens to get selected by the laboratory subsampling tool. Uncontrolled sampling variables are one of the reasons why "duplicate" samples (even from the same jar!) can yield very different analytical results. Particle size effects are closely related to the nugget effect, and they work together produce data variability over and above the location-to-location variability expected for a contaminated site.

Since contaminant concentrations can vary markedly over tiny to large scales, there needs to be a scientific justification for selecting the "right" sampling scale. This choice is straightforward only if the intended use of the data is understood before data are collected. In order for data quality to support environmental decision-making, data representativeness demands that the scale of sample support be chosen to correlate with the scale of the intended decisions (i.e., "decision support"). For example, if the decision is to prove on-going atmospheric deposition of a contaminant, then the sample support must be representative of atmospheric deposition (the first inch or so of surface soil). If there is a concern that lead contaminated dust is blowing from contaminated soil across the site boundary and into people's homes, then the lead concentration in dust sized particles is the correct sample support.

If the intended data use dictates how sampling variables will be controlled, it may be recognized that sometimes high variability at one scale may be inconsequential when viewed at the scale of decision-making. To avoid having variability at the sample scale bias data interpretation at the decision scale, sample collection and processing needs to be structured so that irrelevant variability does not introduce errors into the decision-making process (for example, determining an average for an exposure unit). If small scale variability is important to the decision, then that information must be preserved (for example, the effect of particle size on lead concentrations).

<u>Statistical implications:</u> As stated in the previous section, valid statistical conclusions cannot be supported unless sampling units and their dimensions are chosen in a manner to ensure that variables correlate with the decision unit and its dimensions. Great caution must be exercised when data of unknown sampling quality (i.e, with uncontrolled sampling variables) are fed into statistical calculations. Part of the training to use statistical software must be educating users to critically assess all aspects of data quality before running statistical models or accepting statistical outputs.

Decision-Driven Populations

Classical statistics assumes a single population. However, physical mechanisms, as mentioned above, often cause environmental contamination to have poorly defined populations. It is very important for the project team to define for the project what a "population" is in the context of the decisions to be made. It is common practice to make the unspoken assumption that the area defined by property boundaries is considered to be the statistical population of interest. But this definition is usually dysfunctional for most data used to estimate exposure risk and design cleanup strategies. From the perspective of these activities, a single site will usually contain more than one population. For example, there may be a portion of the site where spills occurred and other portions where there was no spillage. Two populations were created at the time of contaminant release, one with very high concentrations and one with no anthropogenic contamination.

Later human activities (construction, etc.) might cause mixing along the borders of these two populations, so that a third population might be defined based on its intermediate concentrations.

So, in terms of an environmental cleanup, how should a population be defined? The most practical and useful approach is to define a population by simultaneously considering two attributes: matrix volumes with similar characteristics (similar contaminant concentrations, similar release and fate mechanisms, similar matrix properties such particle size, pH, etc.) plus the decision strategy for compliance, risk or remediation. A population should be defined as the matrix unit that will be targeted for the purpose of making a particular decision (ITRC, 2003). Defining populations in relation to the project decisions creates "decision-driven populations."

As a quick example of defining decision-driven populations, suppose a large volume of soil was contaminated by several discrete spills that created relatively isolated "hot spots" within the matrix volume. Therefore, there are soil volumes with high concentrations, volumes with low or no contamination, and a continuum of concentrations between the two extremes. The project goal is to remove and dispose of contaminated soil. One might initially define only two populations: "dirty" and "clean." But additional consideration of what will be done when the "dirty" soil is removed shows that its disposal can take two forms: concentrations below a certain regulatory threshold are considered clean, concentrations higher than a certain value must be incinerated (an expensive disposal option), and concentrations in between the two can be disposed into a landfill (which is less expensive than incineration). A cost-benefit analysis demonstrates that the best strategy is to have the disposal decisions drive definition of three populations: 1) clean soil matrix to remain on site, 2) soil to be disposed into a landfill, and 3) soil that must be incinerated. The sampling design is structured to detect and delineate these decision-driven matrix populations, then segregate them into their three separate bins as they are identified, bounded, and removed.

<u>Statistical implications:</u> There is the notion that statistical assessments are desirable because they are more "objective" and "quantitative." First of all, any appearance of objectivity or quantitative confidence in the output is misleading if inputs to the model are little more than guesses that may in fact be completely inaccurate. Second, recall that the purpose of statistics is to draw inferences about a population when the true state of the population is unknown and samples are relatively few. When a CSM and the concept of decision-driven populations are used to construct sampling plans, there may be little need to use statistics to justify inferences. When coupled with new technologies that support high density sampling, decision-driven populations may be delineated and characterized directly. Sufficiently high densities of data in the right places can be used directly to construct reality, eliminating the need for large extrapolations and their associated uncertainties. Hypothesis testing of site reality (i.e., testing and refining the CSM) through judgmental sampling may be preferable over hypothesis testing of a statistical model which may have little in common with the site reality.

At other times, statistical inference is needed for decision-making, such as when an estimate of the population average, along with the uncertainty of the estimate, is the basis of a decision, such as risk assessment. When statistics are used, applying the concepts in this paper will allow parameter estimates to be more realistic and for the uncertainty in the estimate (often expressed as a confidence interval) to be reduced, increasing the confidence that the decision indicated by the statistical model is actually correct.

References

ITRC (Interstate Technology and Regulatory Counsel). 2003a. Technical Regulatory Guidance for the Triad Approach: A New Paradigm for Environmental Project Management. (http://www.itrcweb.org/Documents/SCM-1.pdf)

ITRC (Interstate Technology and Regulatory Counsel). 2003b. Characterization and Remediation of Soils at Closed Small Arms Firing Ranges. Available on-line at http://www.itrcweb.org/SMART-1.pdf

Steel R G D, J H Torrie, and D A Dickey. 1997. Principles and Procedures of Statistics: a Biometrical Approach, 3rd ed. McGraw-Hill, Boston, MA, USA.

USEPA. 1993. Data Quality Objectives for Superfund. p. 41.